Link Structure based Community Detection 알고리즘의 제안과 소셜 네트워크 분석 및 비주얼라이제이션을 위한 사례 연구

Link Structure based Community Detection for Social Network Analysis and Visualization

온병원, Byung-Won On*, 이인규, Ingyu Lee**, 이만재, Manjai Lee***

요약 커뮤니티 스트럭처는 소셜 네트워크를 이해하는 주요한 특성 가운데 하나이다. 커뮤니티 디텍션에 대한 최근 연구에 따르면 실제 소셜 네트워크에서 중첩된 커뮤니티들을 쉽게 발견할 수 있으며 특정 노드들은 하나 이상의 커뮤니티에 속한다. 기존의 방법론들이 중첩된 커뮤니티들을 효과적으로 찾지 못하는 동안 우리는 이 논문에서 링크 클러스터링 알고리즘을 이용한 새로운 커뮤니티 디텍션 알고리즘을 제안한다. 그리고 우리의 방법론을 소셜 네트워크에 적용함으로써 발견된 커뮤니티 스트럭처를 기반으로 하여 주어진 네트워크에 대한 분석 및 비주얼라이제이션을 시도한다. 이러한 사례 연구는 우리의 제안 방안이 우수함을 간접적으로 증명한다.

Abstract Community structure is one of main characteristics in order to understand social networks. Recent studies on the community detection problem have shown the existence of overlapped communities around real social networks in which some nodes belong to multiple communities. While existing methods seldom find such overlapped communities, in this paper, we propose a novel community detection approach based on the link clustering algorithm. Furthermore, by applying our proposal to a real social network, we study the principles of the given social network by means of community structures found.

핵심어: Overlapped community detection, link clustering, Social network analysis and visualization

본 논문은 2011 년 서울대학교 차세대융합기술연구원 학술 연구비 지원에 의하여 연구되었음

*주저자 : 서울대학교 차세대융합기술연구원 연구교수 e-mail: bwon@snu.ac.kr

**공동저자 : 트로이대학교 경영학부 조교수 e-mail: inlee@troy.edu

***교신저자 : 서울대학교 차세대융합기술연구원 특임연구위원; e-mail: manjai@snu.ac.kr

1. 서론

커뮤니티(community)는 소셜 네트워크에서 a subset of nodes 로 정의된다. 동일 커뮤니티 내에 속한 노드들간에 복잡한 연결을 보이며 다른 커뮤니티 간의 노드 연결은 많지 않은 것이 특징이다. 예로서 웹 상의 한 커뮤니티는 동일한 토픽으로 연결된 웹 페이지들의 묶음을 말한다. 이러한 커뮤니티의 특징을 발견하는 것은 소셜 네트워크의 특성을 이해하고 비주얼라이제이션 하는데 큰 도움을 준다. 더욱이 커뮤니티는 개인 프라이버시를 노출하지 않으면서 네트워크의 특징을 알 수 있게 해주는 장점이 있다. 이러한 이유로 인해서 커뮤니티 디텍션 (community detection)에 대한 연구가 데이터 마이닝 및 소셜 네트워크 소사이어티에서 활발히 진행되고 있다.

커뮤니티 디텍션에 대한 이전 연구들은 주로 주어진 소셜 네트워크를 similarity matrix 으로 변환한다 [3]. 그 매트릭스의 row 와 column 은 두 노드들을 의미하고 그 노드들간의 similarity 값이 매트릭스에 저장된다. 그림 1 은 a similarity matrix 의 예를 보인다. A ~ J 는 노드들을 의미하고 a ~ n 은 링크들을 나타낸다. 그리고 *k*-means, hierarchical clustering 또는 spectral

clustering 알고리즘을 사용하여 노드 간의 similarity 값을 고려하여 각 노드를 단 하나의 커뮤니티에 속하도록 전체

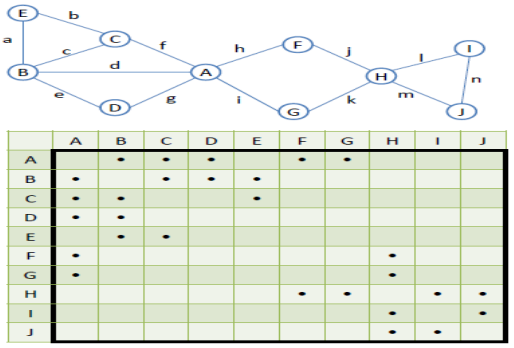

그림 1. 소셜 네트워크와 노드-노드 매트릭스의 예

네트워크를 a set of subgraphs (communities)로 파티션(partition)한다. 그러나 이러한 알고리즘들은 네트워크에서 중첩된 커뮤니티(overlapped communities)들을 찾을 수 없다. 반면에 실제 네트워크에서는 중첩된 커뮤니티를 쉽게 확인할 수 있다. 예를 들면, 협력 네트워크 (collaboration network) 에서 한 과학자(노드)는 데이터마이닝 커뮤니티와 인공지능 커뮤니티에 속한 과학자들과 공동 연구를 수행하는 것이 일반적이다.

이러한 문제를 해결하기 위해, 우리는 기존의 a node-node based similarity matrix 가 아닌 a link-node based similarity matrix 를 사용하는 link clustering 알고리즘을 제안한다. 그림 2 는 링크-노드 매트릭스를 나타낸다. 매트릭스의 row 는 링크를 의미하고 column 은 노드를 나타낸다. 이러한 링크-노드 매트릭스를 사용함으로써 clustering algorithm 은 각 링크 $e_{ij}(v_i, v_j) \in E$ 를 서브그래프 G_k에 어사인(assign)하고 자연스럽게 두 노드 v_i와 v_j 또한 G_k에 어사인 된다. 따라서 v_i 또는 v_j는 여러 개의 다른 커뮤니티들에 속하게 되어 중첩된 커뮤니티들을 발견할 수 있다.

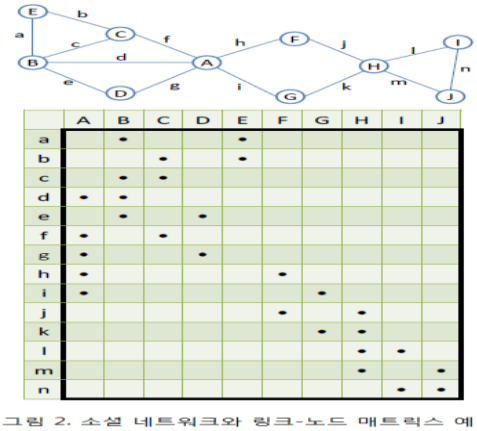

그림 2. 소셜 네트워크와 링크-노드 매트릭스 예

이러한 중첩된 커뮤니티들은 소셜 네트워크를 이해하는데 중요한 도구가 될 수 있다. 따라서 이 논문에서 Link Clustering 알고리즘을 제안하고 2 장에서 자세히 논의한다. 그리고 4 장에서 중첩된 커뮤니티를 사용하여 소셜 네트워크를 분석하는 사례 연구를 수행한다. 타깃 소셜 네트워크로서 우리는 제 18 대 국회의원들과 쟁점이 된 법안들의 데이터를 가지고 소셜 네트워크를 구성한다 [4]. 이러한 데이터를 사용하게 된 주요 동기로는 실생활에서 일반 대중들이 항상 관심을 보이는 주제이고 raw data(예. 국회의원들의 법안투표) 만으로 상황(예. 국회의원들의 투표성향)을 쉽게 이해하기 힘들다. 더욱이 유권자들이 각 의원에 대해 갖는 이미지는 의원들이 눈에 띄는 몇몇 사건이나 이미지 메이킹으로서 결정되고 있는 현실에 비추어볼 때, 국회의원들에 대한 소셜 네트워크 분석을 통해 법안 투표 결과에서 나타난 의원의 객관적 정보를 유권자에게 제공할 수 있게 된다. 그리고 3 장에서 언급한 방법론에 의해 국회의원의 소셜 네트워크를 만들고 노드(국회의원) 간의 관계는 투표성향이 유사한 지를 나타내는 링크로 표현된다. 그리고 링크 클러스터링 알고리즘을 적용하여 국회의원 네트워크에서 중첩된 커뮤니티들을 발견한다. 이미 각 국회의원의 당적을 알고 있기 때문에, 중첩된 커뮤니티에 속한 노드들이 대체로 소신 투표하는 국회의원들이고 중첩되지 않은 커뮤니티에 속한 노드들은 당론을 따르는 투표자일 가능성이 크다. 또한 커뮤니티에 속한 노드들을 분석함으로써 그 커뮤니티가 진보 또는 보수 성향인지를 파악할 수 있다. 따라서 이 논문에서는 국회의원의 법안 투표 결과를 이용하여 소셜 네트워크를 생성하는 방법론을 제안하고 비주얼라이제이션을 수행한다. 또한 중첩된 커뮤니티를 찾기 위해 링크 클러스터링 알고리즘을 제안하고 소셜 네트워크에 적용함으로써 중첩된 커뮤니티에 속하는 노드들을 발견한다. 우리의 가설에 따르면 각 커뮤니티는 진보 또는 보수 성향을 띄며 중첩된 노드들은 당론을 따르기 보다는 소신 투표하는 국회의원들로 4 장에서 심도 있게 논의한다.

2. Link Structure based Community Detection

2.1 링크 클러스터링 알고리즘

입력은 소셜 네트워크로 노드와 링크 (노드 간의 관계)로 구성된다. 우리 실험에서 노드는 국회의원이고 링크는 법안투표성향의 유사함을 -1 ~ +1 값으로 수량화하였고 이 값은 링크의 웨이트 (weight)로 사용된다. 알고리즘의 출력은 서브그래프 셋으로 서브그래프는 소셜 네트워크의 커뮤니티에 해당한다. 기본적으로 링크 클러스터링 알고리즘은 링크-노드 매트릭스를 k-means 클러스터링을 사용하여 전체 네트워크를 서브그래프 셋으로 파티션(partition)한 다음, 커뮤니티 간의 hierarchy 에 의해 유사한 커뮤니티들을 통합하는 과정을 거치면서 최종적으로 중첩 커뮤니티를 발견한다.

알고리즘	링크 클러스터링 (Link Clustering)
Input	Social Network G = (V, E)*
Output	A set of subgraphs $\{G_1, \ldots, G_k\}$
Step 1	// Convert G to a link-node matrix A

	for $e_{ij}(v_i, v_j) \in G$ $A(v_i, v_j) = A(v_j, v_i) = \text{Weight}(e_{ij})$ $A = A \times A^T$
Step 2	// *k*-means Clustering $\{G_1, \dots, G_k, \dots, G_{2k}\}$ = *k*-means(A, 2k)
Step 3	// Assign nodes to the subgraphs for $e_{ij}(v_i, v_j) \in G_k$ Assign v_i and v_j to G_k

Step 4	// Merge the subgraphs
	// Sort $\{G_1, \dots, G_{2k}\}$ by # of nodes per subgraph
	do
	for $G_i \in \{G_1, \dots, G_{2k}\}$
	Merge G_i to G_j such that $G_i \subseteq G_j$
	until # of subgraphs > *k*

*V: a set of nodes & E: a set of links

2.2 *k*-means 클러스터링

k-means 클러스터링은 a set of *d*-dimensional vectors $\{x_1, \dots, x_n\}$ 에 대해 within-cluster 의 square sum 을 최소화하면서 n 벡터들을 k community set $\{C_1, \dots, C_k\}$으로 파티션한다 ($k \leq n$).

$$\arg\min_C \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2 \quad (1)$$

(1)에서 μ_i는 커뮤니티 C_i에 속하는 벡터(노드)들의 평균값(mean)을 의미한다 [2].

2.3 기존 방안과 비교 및 분석

기존 논문들과 비교하여 우리가 제안한 방안은 향상된

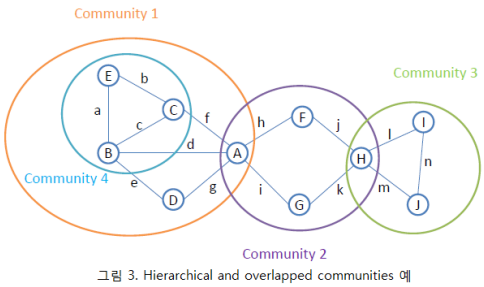

그림 3. Hierarchical and overlapped communities 예

특징을 보이며 보다 정확한 결과를 얻을 수 있다. [1]의 방안은 우리의 제안 방안과 비슷하게 a link-node based similarity matrix 을 사용하여 클러스터링 하였다. 그러나 우리의 링크 클러스터링 알고리즘은 발견된 커뮤니티 간의 hierarchy 관계를 찾아 유사한 커뮤니티들을 머지(merge)함으로써 보다 정확한 결과를 얻을 수 있게 고안되었다. 그림 3 은 중첩된 커뮤니티들과 함께 커뮤니티 간의 hierarchy 을 보인다. 예를 들면, 커뮤니티 4 는 커뮤니티 1 의 child 관계를 보인다.

3. 소셜 네트워크 생성

사례 연구로서 '열려라 국회' 웹사이트로부터 327 명의 18 대 국회의원들의 신상명세(이름과 당적)와 총 63 개의 쟁점이 된 의안 투표 결과를 수집하였다 [5]. 각 의안투표 결과 데이터는 국회의원 성명, 당적, 찬반투표 등을 포함한다. 찬반투표의 경우에는 찬성 / 반대 / 기권 / 불참 / 출장 / 청가 / 결석 등의 투표 행위로 나뉘어진다. 찬반투표를 수량화 하기 위해, 각 국회의원이 찬성했을 경우에는 1, 반대했을 경우에는 -1, 그리고 출장 및 청가에는 0 값을 할당한다. 그러나 기권 / 불참 / 결석의 경우에는 특정 목적을 가진 또 다른 형태의 투표행위이기 때문에 모든 국회의원의 25% 이상이 기권 / 불참 / 결석을 했을 경우에는 법안에 반대한 행위로 간주하여 -1 값을 분배하고 그렇지 않을 경우에는 0 값을 할당한다. 이러한 방식을 통해 각 국회의원의 63 개의 법안 투표 결과를 63 개의 dimension 을 가진 벡터로 표현할 수 있다. 즉 벡터 X={1, 1, 0, 1, …, -1}를 가정하면, 국회의원 X 는 첫 번째와 두 번째 법안에서는 찬성을, 세 번째 투표에서는 출장으로 인해 투표에 참석하지 못했고, 마지막 63 번째 법안에서는 반대표를 행사했음을 알 수 있다.

이러한 327 벡터 셋으로부터 임의의 두 벡터 간의 값들에 대한 상관관계(correlation)가 존재하는 지를 Pearson correlation coefficient 을 통해 측정한다. Pearson 의 상관계수(correlation coefficient) r 의 수식은 다음과 같다.

$$r = \frac{\sum(x_i - \bar{x})(y - \bar{y})}{(n-1)\sigma_x \sigma_y} \quad (2)$$

위의 수식에서 $x \in X$와 $y \in Y$와 같이 두 벡터들이 주어지면 상관계수가 구해진다. x_i 는 벡터 X 의 i-th x 의 스칼라 값이고 $\bar{x}$는 모든 x 의 평균값을 나타낸다. 그리고 n 은 벡터가 포함하는 모든 x 의 수를 의미하고 σ_x는 모든 x 의 표준편차를 말한다. 만일 r 값이 +1 에 가까우면 두 벡터 X 와 Y 는 positive correlation (linear relationship)이 존재한다. 반면에 -1 에 가까우면 두 벡터는 negative correlation 이 존재하며 0 값에 가까울수록 두 벡터 간에 correlation 이 존재하지 않는다 (diffusion). 이와 같이 국회의원 간에 -1 ~ +1 에 해당하는 상관계수 값이 구해지며, 두 국회의원 X 와 Y 의 63 개 법안에 대한 투표성향이 찬성으로 비슷할 때, X 와 Y 의 상관계수는 +1 에 가깝고 반대의 경우에는 X 와 Y 의 상관계수는 -1 에 근접하며 0 값에 가까울수록 X 와 Y 국회의원의 법안 투표성향에 아무런 연관성을 찾을 수 없다.

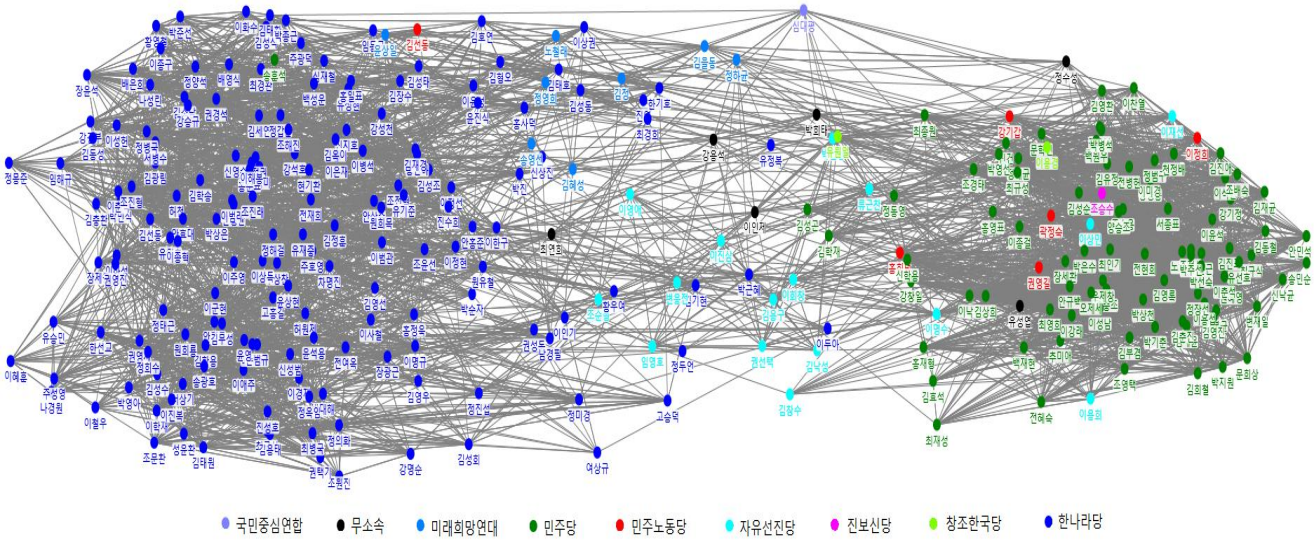

그림 4. 18대 국회의원 소셜 네트워크 비주얼라이제이션

지금까지 327 명의 국회의원들과 63 개의 법안을 사용하여 각 국회의원 간에 투표성향을 Pearson 상관계수로 정의하였다. 이러한 셋팅은 자연스럽게 소셜 네트워크로 변환할 수 있다. 즉 각 국회의원은 소셜 네트워크에서 각 노드를 이루고 두 노드 간의 링크 웨이트(link weight)는 상관계수 값이 된다. 더욱이 이러한 소셜 네트워크는 327×63 매트릭스 (matrix) 형태로 변환된다. 매트릭스의 row 는 각 국회의원 벡터를 의미하고 각 column 은 63 개의 법안 중에 하나를 나타내며 -1 ~ +1 값을 포함한다. 다음, 매트릭스의 각 column 에서 $||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ 을 구하여 나누어줌으로써 노멀라이즈(normalized)하고 2 장에서 제안하였던 링크 클러스터링 알고리즘의 입력으로 사용한다.

4. 소셜 네트워크 분석 및 비주얼라이제이션

3 장에서 논의했던 방법론을 이용하여 국회의원 소셜 네트워크를 생성한 다음, [6]을 사용하여 그림 4 와 같이 비주얼라이제이션 하였다. 그림 4 에서 노드 간에 링크가 존재하면 서로 끌어당기고 링크가 없는 노드 간에는 밀어당기는 알고리즘에 의해 소셜 네트워크가 비주얼라이제이션 된다. 임의의 두 노드 간의 링크의 존재는 법안 투표 성향이 유사함을 나타낸다. 그림 4 에서 보면 한나라당과 민주당의 쟁점법안에 대한 투표성향이 극명하게 나뉘어짐을 알 수 있다. 또한 범여권과 범야권 간의 투표 관계 역시 비슷한 성향을 보인다. 예를 들면 민주노동당 김선동 의원(좌측 상단)을 제외한 모든 민주노동당 당원들의 투표성향은 민주당 소속 국회의원들과 유사하고 두 당의 쟁점법안 투표 성향은 중첩된다. 그러나 예외적으로 민주노동당 김선동 의원의 경우에는 한나라당 국회의원들과 투표 성향이 비슷한 것으로 보이지만, 이는 실제로 김선동 의원과 한나라당 소속 국회의원들과 투표 성향이 비슷한 것이 아니라 김선동 의원이 최근(2011 년 4 월)에 보궐선거로 당선되었기 때문에 대부분의 쟁점 법안 투표에 참가하지 못했고 우연히 최근 법안 투표에서 한나라당과 비슷하게 투표함으로써 나온 결과이다. 이러한 소셜 네트워크에 링크 클러스터링 알고리즘을 적용하여 2 개의 커뮤니티들을 발견하였고 한 커뮤니티는 대부분의 국회의원들이 범여권 소속이었으며 멤버들의 투표 성향을 고려했을 때, 보수 성향의 커뮤니티로 파악되었다. 반면에 다른 커뮤니티는 대부분의 국회의원들이 범야권 소속이었고 진보 성향의 커뮤니티 임을 알 수 있었다. 또한 두 커뮤니티에 중복되어 발견된 국회의원들은 당론을 따르기 보다는 쟁점 법안에 대해 소신 투표 성향을 보이는 국회의원들로 그림 4 의 소셜 네트워크의 중간 영역에 분포한 노드들임을 확인할 수 있었다.

5. 결론

이상과 같이 링크 클러스터링 알고리즘을 제안하여 18 대 국회의원들의 소셜 네트워크를 생성하고 커뮤니티 스트럭처를 기반으로 하여 소셜 네트워크 분석 및 비주얼라이제이션을 수행하였다.

참고문헌

[1] Y. Ahn, J. Bagrow, S. Lehmann, “Link Communities Reveal Multi-scale Complexity in Networks”, Nature, 466(7307), 2011, pp. 761-764.

[2] A. Jain, “Data Clustering: 50 Years Beyond K-Means”, Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD’08), Osaka, Japan, May 20-23, 2008.

[3] M. Newman, “Detecting Community Structure in Networks”, Eur. Phys. J. B(38), 2004, pp. 321-330.

[4] 박한우, “블로그에 나타난 정치인 네트워크: 17 대 국회의원을 대상으로, 한국언론학보, 51 권 3 호, 2007.

[5] “열려라 국회” 웹사이트, http://watch.peoplepower21.org/.

[6] NodeXL, http://nodexl.codeplex.com/.